DOCUMENT RESUME

ED 279 703 TM 870 090

AUTHOR Welch, Wayne W.
TITLE Some Thoughts on the National Assessment of Educational Progress in Science.
PUB DATE 86
NOTE 19p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Testing; Elementary Secondary Education; Evaluation Needs; Federal Programs; *Measurement Objectives; National Surveys; Needs Assessment; Program Attitudes; Research Design; Resource Allocation; *Science Education; *Science Tests; Scientific Attitudes; *Testing Problems; Testing Programs
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

This paper evaluates the procedures and results of the National Assessment of Educational Progress (NAEP) in science. A brief description of the 1981-82 assessment, funded by the National Science Foundation, precedes a discussion of strengths and weaknesses of the current science assessment program, conducted by Educational Testing Service. Major strengths include: high technical quality; accurate portrayals of the national picture; consistent procedures enabling comparison across time; appropriate content; policy impact; and identification of report group differences. Criticisms include: (1) the science assessment cycle has been unnecessarily irregular; (2) assessments tend to focus on traditional subject matter content and to collect little additional information; and (3) subject matter experts who possess a commitment to their field do not seem to have sufficient interest in the data to attempt the difficult task of analyzing unfamiliar data tapes. The assessment program could be strengthened by requiring contractors to include the services of experienced subject matter researchers into the budget proposals. Data should be regularly gathered, and the measurement of attitudes toward science should be improved. Lines of communication between NAEP and the community of educators is essential for continued support of the assessment and for maximizing the educational impact of the information being gathered. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279703

Some Thoughts on the National Assessment of Educational Progress in Science

Wayne W. Welch

University of Minnesota

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W.W. Welch

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

TM 870 090

BEST COPY AVAILABLE

SOME THOUGHTS ON THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS IN SCIENCE

by

Wayne W. Welch
University of Minnesota

Prepared for **The Study Group on National Assessment.** Lamar Alexander, (Chairman), H. Thomas James, (Vice-Chairman). July, 1986.

SOME THOUGHTS ON THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS IN SCIENCE

Wayne W. Welch
University of Minnesota

The National Assessment of Educational Progress (NAEP) was established in the mid-sixties to periodically measure the educational performance of children and young adults. The first subject areas to be assessed were science, writing, and citizenship in 1969-70. Subsequent science assessments were conducted in 1972-73 and in 1976-77. However, in the late 1970s, dwindling financial support for NAEP, declining national interest in science, and a congressional mandate to focus on the "basics" (reading, writing, and mathematics) created a testing schedule that postponed the next science assessment until 1988-89. Some science educators were concerned that a 12-year hiatus between assessments was too long to effectively monitor educational performance in science. With encouragement from a few National Science Foundation (NSF) staff members, I wrote and submitted a proposal to NSF to conduct an interim science assessment in 1981-82. A grant of approximately $500,000 was made to the University of Minnesota to conduct the study with a large sub-contract ($300,000) going to the NAEP contractor at that time, Educational Commission of the States, to "piggy-back" the science testing on the scheduled mathematics and social studies assessment they were conducting that year.

The 1981-82 assessment was in several ways unique from previous assessments. It was funded by NSF, not the National Center for Education Statistics (NCES). It focused on achievement in the science-technology-society realm and contained large numbers of inquiry and attitude items. It was designed to generate data not only for assessment purposes but for science education research as well. As such, more information than usual

was gathered on student and school characteristics (e.g., science anxiety, class enrollments, science encouragement, and computer experience). Furthermore, the grant was made to the University of Minnesota which subcontracted to NAEP for test printing, sample selection, test administration, and item scoring. We at Minnesota were responsible for item selection, writing background questions, data analysis, research studies, and preparation of reports.

This assessment was also different from others in that all achievement items were selected from the prior assessment (1976-77), and we only had one booklet of items for 9-year-olds. These restrictions plus the cooperation of NAEP made it possible to conduct this assessment for about one-third the cost of a typical assessment.

Items were chosen from the four major areas of (a) **Science Content**, the body of scientific knowledge, (b) **Inquiry**, the processes by which this knowledge is derived, (c) **Science, Society, Technology**, the implications of this knowledge for mankind, and (d), **Attitudes**, the beliefs and feelings student have toward science. Approximately equal emphasis was given to each area although the number of items varies due to the generally shorter attitude items. Nearly 400 separate questions were used at ages 13 and 17, and about 100 at age 9. A brief outline of the testing domain is presented in Appendix A.

Approximately 25 per cent of this science assessment was devoted to the gathering of supplementary information on student, home, school, and community characteristics. (See Appendix B). This information was included to make it possible to address the research questions posed in the original proposal. It also provides considerable opportunities for other researchers to conduct secondary analyses of the science assessment data, although, except for the Minnesota group, few have done so.

The 1985-86 science assessment being conducted by Educational Testing Service (ETS), the current contractor, is similar in format and emphasis to the 81-82 assessment although new items have been written, the amount of background information being collected is less, and a full-fledged assessment is being conducted among 9-year-olds. (However, the new assessment policy is to test at grade rather than age levels.)

I've described 1981-82 assessment in some detail to illustrate the general nature of the science testing and to provide some background on my experience with national assessments in science. In addition to being responsible for the 1981-82 assessment, I have been involved in all of the other science assessments, either as an item writer, a member of an advisory committee, or as a reactor to several of the NAEP reports. In addition, my colleagues and I received a grant from NSF to develop procedures for conducting secondary analysis of the 1977-78 mathematics data (Anderson, Welch, and Harris, 1982).

As one can see, I am hardly an unbiased observer of NAEP, particularly in science. However, I think my experience provides a reasonable basis for evaluating the procedures and results of the national assessment. My judgments consist of a set of strengths of the program, things that should be retained in future assessments, followed by a list of limitations or shortcomings, areas where change is needed. Improvement should come from capitalizing on the strengths of the system and eliminating or changing the limitations.

Strengths

Probably the greatest strengths of the national assessment are the technical aspects of design, construction, administration, and processing. Thousands of dollars and hundreds of person-hours are spent in defining

content objectives, writing exercises (NAEPese for items), trial-testing these items, sample selection, exercise administration, quality control, and item scoring. These procedures evolved during the formative years of NAEP, are now well-established, and seem to be consistently implemented year after year. In spite of a few problems now and then, the procedures seem to be as effective as we can make them.

Panels of scientists and educators define the testing domain and exercises are written by experts in the field. Several committees of scientists, science educators, special interest groups, laypersons, and government agencies review all items prior to administration.

A stratified national random sample of counties, schools, and children is carefully drawn and checked in order to make precise estimates of national success levels for each item. Test administration is conducted with the most care of any program I have seen. Professional exercise administrators are used in the schools. Paced audio tapes are used simultaneously as students read the questions. The process insures uniformity of testing conditions and allows all students the same time to answer the questions. Participation and response rates typically exceed 90 percent.

Some observers questioned the content of the first two science assessments claiming several important outcomes of science instruction were omitted. However, the last three science tests have broadly sampled the domain of science. (See Appendix A). They have included substantial numbers of questions on the important outcomes of science instruction and have included items that measure the higher level cognitive skills.

In my opinion, the quality of the test items is high. A recent review by a panel of the National Research Council, while finding fault with a few of the NAEP science items, generally gave them good marks. Given the

tendency to focus on multiple-choice items, a concern of some reviewers, I believe the NAEP items are about as effective as we can make them given the limitations of the English language. Not all experts agree, but none have come forward with any examples when challenged to write an item better than those found in the current assessments.

The technical consistency of the assessments gives us the ability to make accurate comparisons over time. While the results to date have been discouraging, for example, a six percentage point drop among 17-year-olds between the first and fourth assessment, our ability to consistently monitor educational performance is indeed a strength of the science assessment.

Initially, comparisons among various groups of respondents, called "report groups" by NAEP were informative. Regional (southeast U. S. is lower), racial (whites outscore nonwhites), and gender (males score higher than females) differences have important policy implications. Unfortunately, these differences have been so consistent in science, in spite of our attempts to change them, that one wonders about their long term value. Questions have been raised about the ability of our educational system to address these issues. However, given that we have monitored these report group differences for only 16 years, it probably makes sense to continue to do so. However, it is frustrating to find the same differences assessment after assessment.

Another strength of the science assessment is the national publicity that the findings have generated. Each time the assessment has been conducted, headlines across America have carried the story. This is due in large part to the declines in achievement that have been noted. Negative results seem to make good headlines. However, the NAEP results have been a good indicator of the health of science education. The declines have been

influential in creating the recent upsurge of public concern about science. Congress, the National Science Board, state and local education agencies have used the NAEP results to justify an increase in the support of science education. The results have had an important policy impact, one of the major purposes of the national assessment program.

In summary, I see as the major strengths of the science assessment program; (1) high technical quality, (2) accurate portrayals of the national picture, (3) consistent procedures enabling comparison across time, (4) appropriate content, (5) policy impact, and (6) identification of report group differences. In spite of the many strengths of the program, there are several problem areas that need to be addressed. A list of the limitations, as I see them, follows.

Limitations

The science assessment cycle has been unnecessarily irregular. The whims of congressional funding levels and changing priorities in federal agencies should not alter the testing cycle. If the interim assessment had not been funded by NSF in 1981-82, the duration between science assessments would have been three, four, and nine years. The situation has been worse in other subject areas. The period between assessments in a given subject keeps changing. Its similar to conducting a national census on an "every so often" basis rather than on a consistent 10-year cycle.

If national assessments are to be effective and useful for policy makers and researchers, data should be regularly gathered. I would recommend a four or five-year cycle. Five years is frequent enough for policy information, however, four years is appealing because it matches with the different grade levels now being tested. That is, in four years today's seventh graders will be tested as eleventh graders. This permits

us to monitor changes over time for the same cohort group, as well as to compare today's eleventh graders with the eleventh graders of the future.

While this would be a nice addition to the assessment data, I don't believe it is absolutely necessary. It depends, in part, upon the total number of subject areas to be assessed and the attendant response burden on the schools. Three subject areas every four years would make testing possible in 12 areas, however, two areas assessed every five years would be acceptable as well. The major challenge is to determine the cycle and then abide by it.

Most assessments tend to focus narrowly on traditional subject matter content and collect little additional information. Science has been more successful than most in breaking out of this mold, but it has not been easy. Staff people at NAEP and NCES tended to resist attempts to gather new information such as course enrollments, indicators of science anxiety, out of school science experiences, and the like. They preferred the routinized procedures that were used with assessments in other areas. Implementing variations in the already complex process of conducting a national assessment creates problems for the contractor. In addition, the federal review process is a barrier to gathering some data. Reviewers sometimes get nervous about attitude questions. This year, for example, the assessment staff was advised to exclude such items as "Do you think scientists should be given money to study ways to solve population problems?" clearly an important science-society issue. Apparently, someone thought this is a politically sensitive topic. While such censuring has not been a major problem, it together with staff reluctance to change, does tend to create a rather fixed and sometimes sterile test. A major advantage of the recent change of contractors was the opportunity it provided to implement some changes in the system.

Education in the sciences includes much more than the traditional content of physics, chemistry, and biology. And it occurs outside as well as within the classroom. Indicators are needed of the broad range of science learning experiences our children are receiving. The national assessments are at present the only means we have of gathering this important data. We must capitalize on this opportunity even further to more effectively monitor the health of science education.

In some ways, the national assessment seems to lack vision. This excellent data resource is seldom exploited to its fullest. In spite of the attempts of NSF and NCES to stimulate increased use of the data base, few investigators have taken advantage of the opportunity. A two-fold problem exists. The testing agencies with the skills and experience to implement the complicated sampling, administration, weighting, and scoring procedures generally have little inherent interest in the findings or in the questions these data might be able to answer. They are technical experts, not subject matter specialists who have an allegiance to the discipline and a sense of the issues that could be addressed with the data.

A second problem is that the subject matter experts who possess a commitment to their field and know some of the burning issues do not seem to have sufficient interest in the data to attempt the difficult task of analyzing unfamiliar data tapes. Although user tapes are available, my sense is that not many researchers use them. The result is an underused data base which was quite costly to obtain.

The assessment program could be strengthened by requiring contractors to budget into their proposals, experienced subject matter researchers who know the field and are familiar with the data. They would be influential in the design of the assessment and responsible for preparing reports other than the routine change results and report group differences that NAEP

usually publishes. At present, these tasks are carried out by contractor staff people who usually have limited interest in curriculum research.

In some ways, I sense a lack of theoretical perspective, perhaps a national policy on the purpose of the national assessments. Is it merely to measure educational performance on periodic basis, or is there an obligation to build into the assessment some strategies and techniques to help make it possible to explain the results? What theoretical framework guides the selection of areas to be assessed, the frequency of assessment, and the nature of the various assessments? And what responsibility does the assessment have for improving education through research and development.

I searched the Secretary's guiding principles (Bennett, 1985) for answers to these questions. While the principles may be reasonable, they are lacking in rationale or justification. What theoretical, research, or political prospective justifies such principles as encouraging states to be more involved or testing out-of-school seventeen-year-olds? I hope the study group will further develop the theoretical underpinnings of the national assessment program to give it a better sense of purpose.

The science assessment needs improvement in its measurement of attitudes toward science. The 1976-77 assessment was the first to include substantial numbers of attitude items. Most of these have been included in the 81-82 and the 85-86 assessments as well. However, many of the items are not measuring what we hoped and several important areas are being ignored. Seriously neglected is the area sometimes called scientific attitudes, or habits of mind, in contrast to attitudes toward the various components of science. These are ways of thinking about the world that have been used successfully in science, for example, suspending judgment, seeking verification in evidence, assuming effects have causes, and willingness to change one's mind in the light of new information.

It is time to examine carefully, the purpose of the attitude assessment, define the domain more precisely, and develop better measures of the attitudes we wish to assess.

No doubt many reviewers will call for NAEP to expand its test items from the traditional multiple-choice format in an attempt to measure other cognitive skills. While I recognize the limitations of multiple-choice items, there are also problems with other item formats. The recent experience of the International Evaluation of Achievement (IEA) in science with "practical" examinations is a case in point. Students were given materials and asked a series of questions which involved conducting experiments using the materials. It is an appealing idea but the large number of concerns raised by the projects' advisory committee clearly indicates a number of problems with this mode of testing. It is difficult to describe precisely the problem you wish students to solve and difficult to judge the accuracy of their solutions. It would be adviseable to seek feedback from the IEA staff regarding this alternative form of testing before advising future assessments to break out of the multiple-choice mode.

Early assessments included a number of different item types and some exercises required equipment, for example, batteries, wires, and bulbs. Other assessments included open-ended questions. Although these were attractive ideas to science educators, they were time-consuming and very costly. As NAEP gained experience, they tended to limit their items more and more to multiple-choice questions. They felt this format yielded the most information in the limited (45 minutes per student) testing time available.

I would recommend further consideration of different item formats, perhaps as an experimental project funded to evaluate alternative measuring

procedures. It should be done by independent group rather than encumbering the contractor with the added burden. My experience suggests it will be difficult to find techniques for science assessment that accomplish the purpose of NAEP better than multiple-choice questions. However, it is time to try.

I have a few additional observations of the assessment process that properly fall under the category of limitations. I don't have recommendations for improving the situation, but rather would like to sensitize the study group to their presence.

National assessments are expensive to conduct, particularly if they are to be done well. The private, non-profit organizations who have the tools and interest on bidding on proposals are expensive. They must hire staff on soft money with its attendant insecurity, and generally charge very high overhead. Furthermore, if my suggestion to actively involve subject matter experts is followed, the cost will be even higher.

I think the assessment is important enough for Congress to appropriate the money. However, it needs to be continuously and adequately funded to attract experienced and competent professionals. The funding uncertainties of past assessments must to be eliminated so that the responsible organization can concentrate on improving their assessments rather then worrying too much about the next funding cycle.

There is a fairly long turnaround time between the gathering of the data and the reporting of the results, usually about 18 - 24 months. This derives, in part, from the complex nature of the task and the limited budget. It was also related to the several sub-contractors involved in the process, each of whom results in the loss of some central control. I believe the current assessments being conducted by ETS will all be done "in-house". It will be informative to see if the delay time can be

reduced. My guess is that it will not, primarily because ETS is new in the assessment business and they are conducting them in four areas simultaneously.

Our limited science assessment of 1981-82 was able to publish its major report in June, 1983 (Hueftle, Rakow, and Welch, 1983). Testing of 13-year-olds began in October of 1981, however, the 17-year-old testing was not completed until May of 1982. We did not receive the data tapes for age 17 until November, 1982 and it then took us another eight months to analyze the data, write and publish the report. Most assessments will take longer than this. Furthermore, it seems prudent not to start the next assessment until the previous one has been completed. Time is a limitation of the process that seems unavoidable. Perhaps it is not surprising when one considers that a typical assessment may include 500 items and involve the testing of 30,000 students in 700 different schools!

The original NAEP philosophy was to gather data on individual items and report the results accordingly. This philosophy has continued but it is largely ignored by those responsible for preparing reports. It is impossible to report useable results for 100 different items without using some kind of summary statistics. Policy makers and researchers require meaningful scale scores to help them interpret the findings. It is time for a reconsideration of the underlying testing philosophy that focuses on items.

My final comment derives from conversations I had with two knowledgeable colleagues. I asked them what suggestions they had for improving the science assessments. Both stated they felt a lack of communication between NEAP and the research community. They had little knowledge of the current assessment program and sensed a possessive or

secretive attitude about the data. One was interested in obtaining the data but had been unable to do it nor was he able to find out when it might be available. The other had been actively involved in the initial assessment planning and implementation but had not received any information from NAEP in over two years.

I am unable to explain the source of these feelings. It may be the shift of contractors from the midwest to the eastern establishment or their perceptions may be exaggerated. However, the concern reminds us of the importance of establishing lines of communication between NAEP and the community of educators in the nation. These linkages are essential for the continued support of the assessment and to help maximize the educational impact of the information being gathered.

REFERENCES

Anderson, R. E., Welch, W. W., & Harris, L. J. (1982). Methodological considerations in the development of indicators of achievement in the national assessment data. Journal of Educational Measurement, 19(2), 113-124.

Bennett, W. J. (1985). Improving our ability to assess educational progress. Statement by the Secretary of Education. Washington, D.C.: Department of Education.

Hueftle, S. J., Rakow, S. J., & Welch, W. W. (1983). Images of science - A summary of results from the 1981-82 national assessment in science. Minneapolis: University of Minnesota, Minnesota Research and Evaluation Center.

APPENDIX A

Testing Domain

Science Content

Biology

Physical Science

Earth Science

Integrated Topics

Inquiry

Science Processes

Decision-Making

Science-Technology-Society

Persistent Societal Problems

Science and Self

Applied Science-Technology

Attitudes

Science Activities

Attitudes Towards:

- Science Classes
- Science Teachers
- Science Careers
- The Value of Science

Socio-Scientific

- Responsibility

APPENDIX B

Background Information

Background Variables

I. Family/Home Variables
   a. Parents' education
   b. Parents' occupation
   c. Science encouragement

II. Student Characteristics
   a. Race
   b. Sex
   c. Ability (Grades in school)
   d. Homework
   e. TV viewing
   f. Science/mathematics courses taken
   g. Computer experience

III. Student Preferences and Values
   a. Science anxiety
   b. Science interest
   c. Fate control
   d. Self esteem
   e. College graduation
   f. Women in science

IV. School Variables
   a. Science instructional budget
   b. Science enrollments
   c. Teacher qualifications
   d. Public-private school
   e. Science emphasis
   f. Computer facilities
   g. Science/mathematics requirements
   h. racial distribution
   i. Science budget

V. Community Characteristics
   a. Geographical location
   b. Size and type of city
   c. Percent on welfare
   d. Occupational distribution